2014

HLM Behind the Curtain: Unveiling Decisions Behind the Use and Interpretation of HLM in Higher Education Research

Elizabeth Niehaus
University of Nebraska-Lincoln, eniehaus@unl.edu

Corbin M. Campbell
Columbia University

Karen Kurotsuchi Inkelas
University of Virginia

Published in *Research in Higher Education* 55:1 (February 2014), pp 101–122.
doi 10.1007/s11162-013-9306-7

Published online 3 July 2013.

HLM Behind the Curtain: Unveiling Decisions Behind the Use and Interpretation of HLM in Higher Education Research

Elizabeth Niehaus,[1] Corbin M. Campbell,[2] and Karen Kurotsuchi Inkelas[3]

1. University of Nebraska–Lincoln
2. Teachers College, Columbia University, New York
3. University of Virginia, Charlottesville

Abstract
Hierarchical linear modeling (HLM) has become increasingly popular in the higher education literature, but there is significant variability in the current approaches to the conducting and reporting of HLM. The field currently lacks a general consensus around important issues such as the number of levels of analysis that are important to include and how much variance should be accounted for at each level in order for the HLM analysis to have practical significance (Dedrick et al., *Rev Educ Res* 79:69–102, 2009). The purpose of this research is to explore the use of a 3-level HLM model, appropriate contextualizing of results of HLM, and the interpretation of HLM results that resonates with practice. We used an example of a 3-level model from the National Study of Living Learning Programs to highlight the practical issues that arise in the interpretation of HLM within a higher education context.

Keywords: Hierarchical linear modeling (HLM), Statistical methods, Higher education research

Introduction

Hierarchical linear modeling (HLM) has become increasingly popular in the higher education literature (Cheslock and Rios-Aguilar 2008; Di Bartolo et al. 2011). In illustrating this point, Di Bartolo et al. (2011) conducted an extensive literature review of the top three journals in higher education (*Research in Higher Education*, *Review of Higher Education*, and *Journal of Higher Education*) to investigate trends in published quantitative research. They found that the use of HLM had increased from no articles in 1985 to 11% of all quantitative articles in these journals in 2009. The rising popularity of HLM is not surprising, as it is a technique designed for use with nested data and for examining effects at multiple levels. Data in educational research are almost always inherently nested and research questions often reflect interest in understanding the dynamics between levels—students are nested in classrooms, classrooms in schools, schools in districts, districts in states, and so on. The same is true of higher education—college students and faculty are nested in numerous different organizations (e.g., residence halls, student organizations, academic departments, etc.), which are in turn nested in universities. Often, the sampling strategies employed in higher education research emphasize the nested nature of the data by using cluster sampling strategies (e.g., institutions are first sampled and then students within institutions are sampled).

This paper discusses certain methodological issues that arise in the use, reporting, and interpretation of HLM in higher education. Within the higher education literature, there is significant variability in the current approaches to the conducting and reporting of HLM, without a general consensus around important issues such as the number of levels of analysis that are important to include and how much variance should be accounted for at each level in order for the HLM analysis to have practical significance (Dedrick et al. 2009). For example, some scholars in higher education used HLM and denoted practical significance to group level variables when the total variance at the group level ranged from as little as .1% to more than 30% (for example, Hu and Kuh 2003; Kim 2001; Myers 2011; Umbach et al. 2006). Other scholars chose not to use HLM with similar variances reported at the group level (for example, Cox et al. 2010), while still other authors use HLM (or HNLM, hierarchical non-linear modeling) and do not report an intraclass correlation (ICC) or the variance accounted for at the group level (for example, Kim and Conrad 2006).

Additionally, certain prominent scholars have questioned the use of HLM in comparison to more traditional methods of data analysis (e.g., Astin and Denson 2009; Volkwein et al. 2005). These scholars compared estimates in models using ordinary least squares (OLS) regression and HLM, and found that when the proportion of variance explained at the group level is small, the estimates yielded using both methods are very similar. Other scholars have tried to deal with nested data and cluster sampling with econometric multi-level modeling techniques, such as panel models, fixed effects, and random effects models (Cheslock and Rios-Aguilar 2011; Toutkoushian 2001).

In light of these debates and uncertainties in an increasingly popular method of analysis, this paper will focus on certain mechanics of HLM for higher education scholars. The purpose of this research is to explore the use of a 3-level HLM model, the appropriate contextualizing of results of HLM, and the interpretation of HLM results that resonates with practice. We used an example of a 3-level model from the National Study of Living Learning Programs (NSLLP) to highlight the practical issues that arise in the interpretation of HLM within a higher education problem that conceptually fits a 3-level modeling approach. The NSLLP is a perfect example for this study because students are nested within living learning programs, which are nested within institutions—this nesting lends nicely to a 3-level model. While many studies that intend to highlight a methodological issue choose to use simulation data, we purposefully chose a pre-existing dataset in higher education in order to highlight the practical issues that arise in the interpretation of real data in higher education.

Review of the Literature

HLM is a statistical modeling technique that accounts for variation at multiple levels. This technique is particularly useful when data are nested (e.g., students within universities), as is common in educational research, as this nesting violates one of the key assumptions of OLS regression—independence of observations (Raudenbush and Bryk 2002). For example, when students are nested within residence halls, students within the residence hall all experience approximately the same environment. If data analysis only occurs at the individual level, all students within a residence hall are assigned the same residence hall score as if they all independently experienced that environment. This leads to an overrepresentation of the degrees of freedom, a mis-estimation of the standard error, and thus an increased risk of Type 1 errors. By accounting for the nested nature of the data through HLM, researchers can overcome this fundamental problem of OLS regression analysis (Astin and Denson 2009; Dedrick et al. 2009; Raudenbush and Bryk 2002).

According to Raudenbush and Bryk (2002), there are three basic applications of HLM. First, HLM allows for better estimation of individual effects that takes into consideration group-level differences. Second, HLM allows the researcher to partition variance across levels, thus determining how much of the variance can be accounted for at the individual versus group level(s) (the variance accounted for at the group levels is often referred to as the ICC). Finally, HLM allows for exploring cross-level interactions. In other words, the researcher can examine how the effect of a level-1 predictor, such as the effect of race on sense of belonging, might change depending on the environmental context, such as living learning program participation. If such an effect exists, HLM can also allow the researcher to model that interaction. Cheslock and Rios-Aguilar (2011) also noted that HLM allows researchers to estimate level-1 coefficients for groups with smaller sample sizes than would be possible using other forms of analysis, such as linear regression.

Due to the relatively recent growth in use of HLM in higher education research, and in educational research more broadly, the use and reporting of HLM varies widely in the field. According to a recent review of the use of HLM in educational research, Dedrick et al. (2009) found variation in the application and reporting of HLM. As they described of many of the articles reviewed, "one could not determine how many models were estimated, what covariance structure was assumed, what type of centering if any was used, whether the data were consistent with assumptions, whether outliers were present, or how the models were estimated" (p. 69). While there may be many methodological challenges using HLM in the study of higher education, we focus on two questions that have particular implications for practice: the use of 2- versus 3-level models and differences in the interpretation of variance accounted for at each level of analysis.

How Many Levels?

In their review of educational literature, Dedrick et al. (2009) found that most studies using HLM in educational research focused on two levels. This is also the most common use of HLM in the higher education research (Cheslock and Rios-Aguilar 2008), which often focuses on students or faculty within universities (for example, Kim 2001; Kim and Conrad 2006; Myers 2011; Porter 2006; Umbach et al. 2006). While this may be sufficient for some research questions, it is sometimes the case that higher education data are nested within at least three levels. For example, student data may be nested within residence halls within institutions or faculty data may be nested within departments within institutions. There are theoretical reasons why 3-level models may be appropriate in higher education research. For example, Astin's (1991) input-environment-outcomes (I-E-O) model describes both within-institution and between-institution environments. Additionally, early research on college environments suggested that sub-cultures or proximal environments matter just as much, if not more, than institution-level environments (Clark and Trow 1966; Moos 1979; Walsh 1989). This is particularly important when researchers are interested in the effects of programs within universities, including residence halls, leadership or service-learning activities, or living–learning programs (LLPs). Indeed, the effects of such programs can be considered of critical importance, since many of these types of environments have been designated as "high-impact practices," or the elements of the college environment most associated with improved student learning (Kuh 2008).

In addition to the theoretical support for using 3-level models in higher education research, there is empirical evidence that ignoring a level of nesting can have detrimental effects on the results of an HLM analysis. In exploring the effect of excluding a level of nesting in a 3-level model, Moerbeek (2004) found that in balanced designs (where all groups have an equal sample size), there were no substantial problems in power or in coefficient estimation. With unbalanced designs, however (which are far more common in higher education research), she found that "one should expect the estimates and associated standard errors of many, if not all, model parameters to be incorrect if a level of nesting is ignored in the data analysis, and hence the conclusions drawn are not to be trusted" (p. 147). These findings point to the importance of considering the use of 3-level models when three possible levels of analysis exist in the data, as they often do in higher education research.

In recent years, a few higher education researchers have begun using 3-level models to explore research productivity of faculty nested in academic departments, which in turn were nested within broader academic disciplines (Smeby and Try 2005); the effect of instructor feedback on student evaluations, considering students nested within courses nested within instructors (Dresel and Rinderman 2011); and in-state graduate out-migration at the student, institution, and state levels (Ishitani 2011). In one of the few studies to use a 3-level model to explore student outcomes, Kugelmass and Ready (2011) examined racial differences in cognitive development of students, taking into consideration differences at the individual, department, and university levels.

While arguments can be made for the use of 3-level models in much of higher education research, there is also a broader argument in the field over whether HLM is necessary or provides any advantage over the traditional, one-level OLS regression approach. For example, Astin and Denson (2009) compared OLS regression and HLM in the analysis of a multi-institutional survey of college students. In a side-by-side comparison of the results of each analysis, they found that the *p* values for institution-level predictors were somewhat higher in the HLM than the OLS, but this did not actually lead to a difference in the findings of significance; *p* values for individual-level predictors were very similar. Astin and Denson argued that while HLM provides a more conservative estimate of institution-level effects, "unless the college effects investigator is interested in whether the slopes of the individual-level predictors vary across institutional units (cross-level effects), it is not clear whether HLM offers any other advantages over OLS regression" (p. 357). They also pointed to the advantage of step-wise OLS regression in allowing the researcher to examine changes in coefficients and the variance accounted for in the model over a series of hierarchical regression analyses. In a study of faculty views of racial diversity, Park and Denson (2009) chose to use OLS regression for exactly this reason, but used a lower *p* value cut-off (.001) for institution-level variables in order to account for the increased Type 1 error rate resulting from nested data.

Another challenge in using HLM is that a 2- or 3-level model requires a larger sample size than might be available for some research questions. For example, in a simulation study using a 2-level model, Maas and Hox (2005) suggested that it would be necessary to have at least 50 cases at level-2 in order to achieve unbiased standard errors. Although 50 programs or institutions may be readily available in large, national datasets, this may be beyond many smaller studies.

How Much Group-Level Variance is Enough?

In addition to the debate about the appropriate level or levels of analysis, another significant issue in the higher education HLM literature is the widely disparate interpretations of how much variance should be accounted for at the group level(s) in order for the HLM analysis to be practically meaningful. A number of published research articles only note the amount of variance explained at each level (i.e., the ICC) in a table (e.g., Umbach et al. 2006) or in a description of the variables used (e.g., Porter 2006), but do not contextualize the results in terms of the variance accounted for at each level. When the ICC is reported, interpretations of what number is practically significant and how to proceed can vary widely. For example, Kinzie et al. (2007) compared engagement of women students at women's and coeducational colleges, and used HLM when they found that the institution level explained between 2 and 12% of the variance in the different outcome measures. Similarly, Hu and Kuh (2003) explored predictors of students' learning productivity and reported an institution-level variance (ICC) of 8.5%. Pike et al. (2011) used HLM to determine the relationship among student and institutional characteristics (e.g., student race and institutional expenditures), student engagement measures, and student cognitive outcomes, and found 2% of the variance explained at the institution level. At the far ends of this spectrum, Dresel and Rinderman (2011) used a 3-level HLM to examine differences in student evaluations of counseling instructors and found that 32.9% of the variance could be explained at the course-level (level-3) and 32.6% at the instructor-level; Myers (2011) explored the relationship between union membership and faculty job satisfaction and found that only .1% of the variance could be accounted for at the institution level. Despite this, Myers still chose to use HLM, explaining:

> When the institutional variance component approaches 0.0 the HLM model essentially reduces to an OLS model. However, there are still strong reasons to employ HLM techniques. In particular, HLM explicitly models the dependency between observations, produces unbiased standard errors, and produces more stable slope and intercept estimates. (p. 669)

While the above-cited researchers all chose to use HLM with ICCs between .1% and 33%, other researchers have made different decisions with ICCs within that same range. For example, Cox et al. (2010) found that 3% of the variance in type and frequency of students' interactions with faculty members was explained at the institution level and chose not to use HLM. As they explained:

> Even if we fit the perfect level 2 model, we would explain no more than 3% of the variability in the frequency of faculty-reported instances of contact with students outside the classroom. Instead, over 97% of the total variance occurs at the individual level, regardless of their institutional affiliation. With this finding in mind—dramatic within-college variance combined with little between-college variance—we chose to focus our

> analysis on the manner in which the behaviors of individual faculty-members shaped their level of interaction with students outside of class. (p. 778)

Steinberg et al. (2009) took a slightly different approach when they examined the prevalence of Pell Grant recipients at colleges and universities (nested within states). They found little evidence of significant variance accounted for at the state-level (although they did not report how much variance was accounted for), so chose to use OLS regression with adjusted standard errors. As there was some state-level variability, they also used HLM to check their OLS regression results. Although they did not report the results of the HLM analysis, they offered to make the results available upon request and stated that they were "virtually identical to those reported" (p. 247) from the regression analysis. In a study of the relationship between institutional selectivity and educational practices, Pascarella et al. (2006) similarly chose to use OLS regression due to low ICCs, but double-checked the results with HLM.

In addition to discrepancies in the use of HLM depending on the amount of variance accounted for at the group level(s), there is also substantial variation in how this variance is reported in research articles. As explained above, some articles do not report this information at all, while others report it in different ways. The most common way is to provide the percentage of variance accounted for at each level or the ICC. However, there are exceptions. In comparing women students' desire to influence social conditions at women's and coeducational colleges, Kim (2001) described the group-level variance in terms of the relative standard deviation—"the institution-level standard deviation (0.17) is small in relation to the total standard deviation (0.76). Given that the outcome of interest was measured on a 4-point scale, the amount of institution-level variation is fairly small" (p. 304). However, she went on to use HLM and even to model the institution-level predictors.

How Important is the ICC for Contextualizing Implications?

The ICC is a particularly important consideration in interpreting HLM results as it allows readers to put significant findings in the appropriate context—if only 2% of the overall variance is at the group level, significant group-level predictors can explain *at most* 2% of the variance in the outcome, leading to questions of practical significance. While researchers using OLS regression can use the easily interpreted R^2 statistic to describe the proportion of variance in an outcome that can be explained by a variable or group of variables in a model, calculating an R^2 or ΔR^2 in HLM does not have the same interpretation (Luke 2004). In HLM determining the proportion of variance explained by each predictor within each level is much more complex (McCoach and Black 2008; Snijders and Bosker 1994), contributing to the importance of considering the ICC at each level of analysis.

While to some extent the differences in judgment regarding the use of HLM or regression may be justified due to differences between what might be practically meaningful in one situation compared to another, the amount of variance being explained can have significant policy implications. Often, researchers who report very small variances at the group levels, or those who do not report at all, still discuss policy implications for statistically significant predictors at these levels. For example, in exploring the difference between graduation rates of African American students at HBCUs and HWCUs, Kim and Conrad (2006) stated that "differences in the college characteristics explain the majority of the college-level variance," (p. 416) but since they state that there is no equivalent to the ICC in HNLM, they never report how much variance is can be explained at the college level.

Similarly, Kim (2001) reported that little variation in women students' desire to influence social conditions existed at the institution level, but then went on to model that variation with five predictors. She reported that "these five variables explain 67% of the total institution-level variation" (p. 305), which is a high percentage of a very small slice of the overall variation in this outcome. Despite this, she concluded, "Attending women-only colleges appears to be more beneficial in developing students' desire to influence social conditions than attending coeducational institution" (p. 308), and "Women-only colleges demonstrated their institutional effectiveness by their positive contribution to students' desire to influence social conditions." (p. 311).

In another example, Pike et al. (2011) explored the relationships among student and institutional characteristics (e.g., student race and institutional expenditures), student engagement measures, and student cognitive outcomes. In this study, Pike and colleagues found that there was a statistically significant variation in cognitive gains across institutions, yet the amount of variation explained at the institution-level was 2%. While student engagement measures accounted for 51% of ICC, that amounted to just over 1% of the overall variance in student cognitive gains.

While it is possible to calculate whether there is a statistically significant amount of variance at the group level(s) (e.g., is the amount significantly different than zero?), there is little guidance as to how much variance is of practical significance. Raudenbush and Bryk (2002) stated that even very small amounts of variation at the group level (e.g., as little as 2–3%) could be important to model, but ultimately this is a judgment call left up to the individual researchers. Contributing to this problem is the lack of agreement in the broader HLM literature about how to report effect sizes that constitute practical significance. Certain scholars in educational measurement and statistics, such as Hedges (2007), have called for the use of specialized effect sizes in multilevel designs that take into account within and between group variation. Other scholars have discussed the use of model fit indices, such as reduction in deviance (Luke 2004) or proportional reduction in variance (McCoach and Black 2008), which both compare models with different predictors to determine whether predictors are significant contributors to the model. Yet other scholars describe multiple methods for effect sizes and how the calculation of effect sizes in HLM is complex and unclear (Roberts and Monaco 2006). Partly because of the lack of agreement about how to denote practical significance, very few articles in the higher education literature report such effect sizes regarding proportion of variance explained at each level for specific predictors.

In light of the recent increase in higher education scholars who use HLM, this paper focuses on two analytic and interpretive concerns in the use of HLM in higher education research: the number of levels and the proportion of variance explained at each level. Although there are many other issues to consider when employing HLM, such as centering of variables, model building, number of cases needed at each level, model fit indices, etc. (Cheslock and Rios-Aguilar 2008, 2011; O'Connell and McCoach 2008), we zero in on these two concerns because they have not yet been addressed in a higher education context and have important implications to practice in using HLM in higher education research. As such, the purpose of this study is to:

(1) Explore the use of a three level model, using the example of data from a national preexisting study in higher education;
(2) Provide an example of reporting of decisions made in the use of HLM, allowing read-

ers to see the judgments made by researchers about proportion of variance at each level; and

(3) Describe the different possible interpretations of results, and the practical implications of the researcher's different decisions to the fundamental questions regarding variance explained and the use of HLM.

Methods

Background of the Study

While the purpose of this study is to explore a methodological question, the study itself still must have a meaningful context. In this case we chose the outcome variable of students' sense of belonging due to its proven importance in the higher education literature (e.g., Nora and Cabrera 1993; Hausmann et al. 2007 Johnson et al. 2007), and the existence of a previous study by Johnson et al. (2007) that explored racial differences in sense of belonging within the context of LLPs. This provides a perfect context for exploring the use of HLM as a method in higher education research. Students are nested in LLPs, which are in turn nested in colleges and universities. Johnson et al. (2007) found significant racial differences in sense of belonging, but did not find any differences in sense of belonging between students who did and did not participate in LLPs. Using linear regression, they were unable to explore whether or not racial differences in sense of belonging differed across institutions or LLPs, or whether the average sense of belonging within LLPs or institutions varied.

Data

The question of whether LLPs vary in terms of students' average sense of belonging or the gap in sense of belonging between White students and Students of Color is a perfect question to be answered by HLM, as HLM allows the researcher to explore both individual and group-level predictors simultaneously. Building on Johnson et al. (2007) work, the current study used data from the 2007 NSLLP, a multi-institutional survey designed to explore the effect of LLPs on undergraduate students. The 2007 NSLLP was administered via a web-based survey with an overall 20.9% response rate, resulting in 22,519 total respondents from 49 campuses. All LLPs at participating institutions were included in the survey. For smaller programs, all students participating in LLPs were invited to participate; for larger programs, a random sample of students received an invitation to participate in the online survey (Inkelas 2007).

In order to provide a practical example of a 3-level model, we decided to build off of a previous study that included constructs at multiple levels but did not use HLM. We built off of Johnson et al. (2007) work, which found that student's sense of belonging was influenced by their race (individual-level variable) but not whether they participated in an LLP (which they modeled on the individual level). As a result, we used HLM to determine whether a student's sense of belonging and whether the racial gap in sense of belonging (i.e., slopes as outcomes) was influenced by individual level (level-1) and LLP level (level-2) predictors, while modeling variance at levels 1 (student), 2 (LLP), and 3 (institution). We used the following variables from the NSLLP dataset and IPEDS (see Table 1 for further information and descriptive statistics):

Table 1. Variables and descriptive statistics

Variable Name	Coding	N	Mean	SD	Min	Max
Sense of belonging	n/a	6,667	12.82	2.45	4	16
Race	0 = White, 1 = Students of Color (African American, Asian/Asian American, American Indian, Hispanic, Multiracial, Other)	6,667	.23	.42	0	1
Gender	0 = Female, 1 = male	6,667	.34	.47	0	1
High school grades	n/a	6,667	11.66	4.19	1	21
SES	n/a	6,667	1.65	.78	1	6
Course-related faculty interactions	n/a	6,667	7.82	2.56	4	16
Residence hall academic climate	n/a	6,667	11.04	2.39	4	16
Residence hall social climate	n/a	6,667	20.77	3.8	7	28
Positive peer diversity interactions	n/a	6,667	11.43	3.03	3	18
Intended co-curricular involvement	n/a	6,667	12.99	3.31	3	18
Mean residence hall academic climate	n/a	386	3.38	.73	1	4
Mean residence hall social climate	n/a	386	1.51	.5	1	2
Sense of belonging as an important outcome	n/a	386	10.74	1.36	5	16
Selectivity of the LLP	n/a	386	20.7	2.12	13	28
Institution size 1000–4999	Dummy coded with referent group of over 20,000	45	0	0	0	0
Institution size 5000–9999		45	.04	.21	0	1
Institution size 10,000–19,999		45	.11	.32	0	1
Control	1 = Private not-for profit, 0 = public	45	.27	.45	0	1
Percent White	n/a	45	.2	.4	0	1
Selectivity	n/a	45	65.18	14.16	27	88
Student/faculty ratio	n/a	45	67.44	15.46	33	99

All variables were entered into the HLM analysis grand-mean centered, except for race which was group-mean centered (as it was of particular interest how the intercepts and slopes for this variable might vary across LLPs and institutions)

- Outcomes: sense of belonging (individual level), gap in sense of belonging between White students and Students of Color (slopes as outcomes)
- Level-1 predictors: gender, socioeconomic status, high school grades, course-related interactions with faculty, residence hall social and academic climates, academic and social transition to college, positive peer diversity interactions, and intended co-curricular involvement in college

- Level-2 predictors: whether or not sense of belonging was a stated outcome of the LLP, whether or not the LLP was selective, and the overall perception of the residence hall academic and social climate within the LLP (mean aggregated from the student level)
- Level-3 predictors: institution size (dummy coded with over 20,000 students as the referent group), institutional control (public vs. private), the percentage of students at the institution who identify as White, institutional selectivity (measured by the percentage of students admitted), and the student/faculty ration.

Data Analysis

As recommended by Raudenbush and Bryk (2002), we first ran a random-effects ANOVA model (in HLM 7.0, using full maximum likelihood estimation) in order to determine the partitioning of variance among the three levels of analysis. The fully unconditional HLM model is represented below:

Level-1: $Y = \pi_0 + e$
Level-2: $\pi_0 = \beta_{00} + r_0$
Level-3: $\beta_{00} = \gamma_{00} + u_{00}$

This model indicates that a student's sense of belonging is a function of the mean sense of belonging in the LLP plus some individual variation. The mean sense of belonging in the LLP is a function of the mean sense of belonging across all LLPs in the university plus some amount of variation between programs, and the mean sense of belonging in the university is a function of the mean sense of belonging for all universities in the sample plus some amount of variation between universities. The random effects for the intercept at level-2 (r_0) and level-3 (u_{00}) are the extent to which mean sense of belonging varies between LLPs and universities, respectively.

After partitioning the variance among the three levels of analysis, we then entered students' race (0 = White, 1 = Students of Color) to the model at level-1, group-mean centered (as one focus was to explore how the intercept of sense of belonging might vary across LLPs), allowing the intercepts and slopes to vary at level-2 and level-3:

Level-1: $Y = \pi_0 + \pi_1(race) + e$
Level-2: $\pi_0 = \beta_{00} + r_0$
$\pi_1 = \beta_{10} + r_1$
Level-3: $\beta_{00} = \gamma_{00} + u_{00}$
$\beta_{10} = \gamma_{10} + u_{10}$

The addition of race to this model at level-1, and to the random effects of race at level-2 and level-3, indicates that a students' sense of belonging is a function of the mean sense of belonging in the LLP, plus some effect of race, plus some individual variation. As race is coded dichotomously and the variable is group-mean centered, the coefficient for the effect of race is the mean difference in sense of belonging between White students and Students of Color. Including the random effect for race at level-2 (r_1) and level-3 (u_{10}) allows the coefficient for race to vary between LLPs and universities; if these random effects are significant, the mean difference in sense of belonging between White students and Students of Color is significantly different in different LLPs and/or universities. Once we had determined whether or not the intercepts (the average sense of belonging) or slopes (the difference in sense of belonging between White students and Students of Color) varied across LLPs or universities, we then entered predictor variables into the model, grand mean centered, following Johnson et al. (2007) previous analysis of this topic (with a few differences based on differences in the 2004 and 2007 administrations of the NSLLP survey).

Next, we ran three different analyses in order to demonstrate impact of the different possible interpretations of the variance accounted for at LLP and institution levels. The first analysis was a 3-level HLM model with predictors at levels 1, 2, and 3. The second analysis was a hierarchical OLS regression, with three blocks of variables (inputs, bridge measures, and environments), and the third analysis was a 3-level HLM with predictors only at level 1. Both the second and third analyses only included the level-1 variables.

The HLM analysis is based only on students involved in LLPs, and only those LLPs that had both White students and Students of Color represented in the sample. Due to missing data, 6,667 students, 386 LLPs, and 45 universities were included in the final HLM analyses with predictor variables. Overall there was a range of 13–653 students per institution, 2–653 students per LLP, and 2–33 LLPs per institution. The assumptions of both the regression and HLM analysis (e.g., normal distribution of error terms and independence of predictor variables) were tested and found to be tenable.

Results

The partitioning of variance through the fully unconditional HLM model found that 96.6% of the variance was explained at level-1, 1.4% explained at level-2 (an ICC of .014), and 2.0% explained at level-3 (an ICC of .020). The variance components at all levels were significant ($p < .001$). This means that of all of the variance in sense of belonging across all students in all programs at all universities, 2.0% of that variance is between universities, 1.4% is within universities and between LLPs, and 96.6% is within LLPs (between individual students).

Adding race to the model at level-1, we found that race was a significant predictor of sense of belonging overall: Students of Color on average reported lower sense of belonging than did White students ($\pi_1 = -.452$, standard error = .0759, t-ratio = −5.964, df = 46, $p < .001$). There was also significant variation in the mean sense of belonging across LLPs (r_0: SD = .258, variance component = .067, df = 276, Chi squared = 382.327, $p < .001$) and across institutions (u_{00}: SD = .396, variance component = .157, df = 45, Chi square = 185.153, $p < .001$). There was *not*, however, significant variation in the *difference* in sense of belonging between White students and Students of Color across LLPs (r_{01}: SD = .307, variance component = .094, df = 276, Chi square = 308.899, $p = .084$) or across universities (u_{01}: SD = .159, variance component = .025, df = 45, Chi square = 52.656, $p = .202$).

At this point we could choose to proceed in a number of different ways, all of which we believe reflect similar decisions made in published research in higher education. Below, we outline three possible decisions about how to proceed with the analysis based on the variance accounted for at the group and institution levels:

1. *Use It*: Employ a full, 3-level HLM model with predictors at levels-1, 2, & 3.
2. *Lose It*: Use OLS regression (thus only including level-1 predictors).
3. *Use It Cautiously*: Use a 3-level HLM model, but only include predictors at level-1.

In the following three sections we describe the rational for each approach and the subsequent results and interpretations based on the three different decisions. In essence, in the following section, we are "trying on" each of three different ways of thinking about the variance explained at each level, making three different assumptions regarding what is practically significant and the importance of statistical significance, in order to explore how each decision may lead to different results and/or interpretation of results.

Rationale for Decision 1: *Use It*

After the fully unconditional model and the model including race as a level-1 predictor were tested, the amount of variance explained at the program and institution level was statistically significant, indicating that HLM is an appropriate method to use to analyze these data. While the majority of the variance was accounted for at the individual level (96.6%), many studies have employed HLM when the variance accounted for at the group level(s) was relatively small (e.g., Kim 2001; Myers 2011; Pike et al. 2011). Although the difference in sense of belonging between White students and Students of Color (the variance component for the slope for race) did not vary across LLPs, the mean sense of belonging did vary significantly across LLPs and institutions. This indicates that there is something that these programs and institutions are doing that can affect sense of belonging overall. As there was significant variance at both group levels, and research on college environments points to the importance of both institution-level and within-institution contexts (e.g., (Clark and Trow 1966; Moos 1979; Walsh 1989), a 3-level model best accounts for the theoretical, statistical, and practically significant variation in sense of belonging and will provide the most comprehensive information to influence policy.

As such, we continued with the HLM analysis by adding predictors at level 1, 2 and 3, following Johnson et al. (2007) analysis of this topic (see Table 1 for a list of variables and descriptive statistics). The results of this analysis showed that at the individual level, race and high school grades were both negative predictors of sense of belonging, while a smooth academic and social transition to college, students' individual experience of the residence hall academic and social climate (measured at the student level), and positive peer diversity interactions were all significant positive predictors of sense of belonging ($p < .001$ for all significant predictors). At the program level, the overall perception of the residence hall social climate within the LLP (mean aggregated from the student level) was a significant predictor of sense of belonging ($p < .001$). At the institution level, only the percentage of White students at the institution was a positive predictor of sense of belonging ($p < .01$) (see Table 2).

Rationale for Decision 2: *Lose It*

After the fully unconditional model was tested, the amount of variance explained at the program and institution level was statistically significant. However, the vast majority (over 96%) of the variance in sense of belonging was between individuals. Such a small proportion of variance accounted for at the group (1.4%) and institution (2.0%) may not be not practically meaningful. Previous research has shown that parameter estimates will be similar between OLS and HLM when group level variance is small (Astin and Denson 2009). As Cox et al. (2010) explained of their decision to use OLS regression instead of HLM when only 3% of the variance could be explained at the group level, even if we were to perfectly model predictors of the level 2 and 3 variance, we would be explaining such a small amount of the overall variance as to be prac-

Table 2. Predictors of sense of belonging using a 3-level HLM with predictors at all three levels

Variable	Coefficient	Standard error	p value
Level 1			
Race (0 = White, 1 = Students of Color)	−.427	.068	<.001
Gender (0 = female, 1 = male)	.061	.056	.276
SES	−.007	.006	.281
High school grades	−.140	.035	<.001
Academic transition to college	.070	.010	<.001
Social transition to college	.185	.009	<.001
Course-related faculty interactions	−.009	.011	.418
Residence hall academic climate	.094	.016	<.001
Residence hall social climate	.138	.010	<.001
Positive peer diversity interactions	.031	.006	<.001
Intended co-curricular involvement	.014	.016	.345
Level 2			
Mean residence hall academic climate	.047	.051	.357
Mean residence hall social climate	−.111	.031	<.001
Sense of belonging as an important outcome	.068	.057	.234
Selectivity of the LLP	.101	.077	.190
Level 3			
1,000–4,999 (vs. over 20,000)	−.659	.332	.055
5,000–9,999 (vs. over 20,000)	−.137	.192	.482
10,000–19,999 (vs. over 20,000)	−.075	.112	.511
Control (1 = private not-for profit, 0 = public)	.101	.172	.559
Percent White	.013	.004	.004
Selectivity	−.007	.004	.119
Student/faculty ratio	.016	.014	.274

Variance components

	Standard deviation	Variance component	df	χ^2	p Value
Level-1 and level-2 (r_0)	.250	.062	337	476.157	<.001
Level-3 (u_{00})	.157	.025	37	72.640	<.001

Deviance = 28820.805749
Number of estimated parameters = 26

tically insignificant. Additionally, using HLM can provide unduly conservative parameter estimates (i.e., Type II error) because of partitioning of variance across groups. Following the lead of previous researchers (e.g., Cox et al. 2010; Park and Denson 2009; Pascarella et al. 2006), we proceeded with a single-level analysis, allowing the most important focus on individual-level predictors of sense of belonging to take center stage in the analysis. A single-level OLS model provided the most parsimonious and practically significant information to guide policy.

The results of this analysis showed that after adding all variables to the model, race ($p < .001$), high school grades ($p < .001$), and course-related faculty interactions ($p < .05$) are all negative predictors of sense of belonging, while a smooth academic ($p < .001$) and social ($p < .001$) tran-

Table 3. Predictors of Sense of Belonging Using Hierarchical OLS Regression

	Block 1			Block 2			Block 3		
	Beta	SE	*p*	Beta	SE	*p*	Beta	SE	*p*
Race (0 = White, 1 = Students of Color)	−.452	.066	<.001	−.389	.060	<.001	−.421	.059	<.001
Gender (0 = female, 1 = male)	.071	.056	.205	.004	.051	.935	.091	.049	.063
SES	.014	.007	.036	.005	.006	.431	.000	.006	.983
High school grades				−.219	.030	<.001	−.163	.029	<.001
Academic transition to college				.094	.009	<.001	.070	.009	<.001
Social transition to college				.266	.008	<.001	.191	.009	<.001
Course-related faculty interactions							−.021	.010	.036
Residence hall academic climate							.102	.014	<.001
Residence hall social climate							.126	.009	<.001
Positive peer diversity interactions							.027	.005	<.001
Intended co-curricular involvement							.029	.014	.039
R^2	.008			.191			.262		
ΔR^2				.183			.071		
p				<.001			<.001		

The regression analysis included 7,780 students, 477 living–learning programs, and 47 institutions

sition to college, residence hall academic ($p < .001$) and social ($p < .001$) climate, positive peer diversity interactions ($p < .001$), and intended co-curricular involvement ($p < .05$) were all positive predictors of sense of belonging. Additionally, we can see that adding the block of bridge variables (as Astin 1991, described, variables "measured at the time the student enters college (input) [that] also signify environmental experiences that can continue to affect the student's development during the college years" (p. 74); in this case, high school grades and transition to college) accounted for an additional 18.3% of the variance in sense of belonging ($p < .001$), while adding the college environments (faculty interactions, residence hall climate, peer diversity interactions, and intended co-curricular involvements) accounted for an additional 7.1% of the variance ($p < .001$). The model overall accounted for 26.2% of the variance in students' sense of belonging (see Table 3).

Rationale for Decision 3*: Use It Cautiously*

After the fully unconditional model was tested, the amount of variance explained at the program and institution level was significant, indicating significant clustering effects. Therefore, as described above, a 3-level HLM is the best method to estimate results and account for nesting of data. However, a vast majority (96.6%) of the variance in sense of belonging was explained at the level of individual students. Only 1.4% was explained by differences across LLPs, and 2.0% was explained by differences across institutions. These levels, while statistically significant, do not represent practical significance that would be substantive for policy implications. If LLP or institution level variables were to be included in the model, they could only explain

Table 4. Predictors of sense of belonging using a 3-level HLM with predictors only at level-1

Variable	Coefficient	Standard error	*p* Value
Level 1			
Race (0 = White, 1 = Students of Color)	−.426	.068	<.001
Gender (0 = female, 1 = male)	.063	.056	.266
SES	−.007	.006	.267
High school grades	−.140	.035	<.001
Academic transition to college	.069	.010	<.001
Social transition to college	.186	.009	<.001
Course-related faculty interactions	−.009	.011	.382
Residence hall academic climate	.100	.015	<.001
Residence hall social climate	.129	.010	<.001
Positive peer diversity interactions	.028	.006	<.001
Intended co-curricular involvement	.013	.015	.374

Variance components	Standard deviation	Variance component	*df*	χ^2	*p* Value
Level-1 and level-2 (r_0)	.313	.098	341	508.147	<.001
Level-3 (u_{00})	.244	.060	44	93.042	<.001

Deviance = 28853.283032
Number of estimated parameters = 15

a maximum of 1.4 and 2.0%, respectively, of the total variance in a student's sense of belonging. However, as Myers (2011) explained, using HLM in this situation still presents advantages over OLS regression in that "HLM explicitly models the dependency between observations, produces unbiased standard errors, and produces more stable slope and intercept estimates" (p. 669). As a result, while we chose to use HLM, we only modeled and interpret variables at the individual level. LLPs and institutions may be able to exert small differences on an individual student's sense of belonging. However, a student's sense of belonging seems to be tied much more closely to characteristics of or actions taken by that individual student.

The results of this analysis show that race ($p < .001$) and high school grades ($p < .001$) are both significant negative predictors of students' sense of belonging, while a smooth academic ($p < .001$) and social ($p < .001$) transition to college, the residence hall academic ($p < .001$) and social ($p < .001$) environment, and positive peer diversity interactions ($p < .001$) are all significant positive predictors of sense of belonging (see Table 4).

Discussion

By looking across the results of the three models, we can see the possible differences in results based on modeling decisions (Table 5). The results obtained through each analytic decision above overall lead to very similar results, with a few key differences. In fact, between the two HLM analyses (including level-2 and level-3 predictors versus only level-1 predictors), the same student level predictors turned out to be significant, and the coefficients were quite similar. This might be different if there were more significant level-2 or level-3 predictors in the model or if more variance was accounted for at either grouping level, but with such a small ICC at levels-2 and 3 it is not surprising that including the level-2 and level-3 predictors did not change the level-1 model substantially.

At the program level, however, there was one key difference between the two analyses. When level-2 and level-3 predictors were included in the model, the overall perception of the residence hall social climate within the LLP was a significant, negative predictor of students' sense of belonging, and the percentage of students who identified as White at the institution was a significant, positive predictor. When the higher level predictors are not included, these statistically significant findings are absent. Yet, this may actually be a better reflection of the broader effects on sense of belonging, as only 1.4% of the overall variation in sense of belonging can possibly be explained by level-2 predictors, and only 2.0% by the level-3 predictors (and it is likely that the predictors in the model only explain a fraction of the overall variance at each of those levels). Discussing the implications of variables that can only predict such a small amount of variance may not be of practical value.

A direct comparison of the two HLM models can be made using the deviance statistic for each model. According to Luke (2004), when one HLM model is a subset of the other, the two models can be compared using the change in deviance, which follows a Chi squared distribution with the number of degrees of freedom being the change in the number of predictors. In doing this, we find that the model with predictors at all levels *does* produce significantly better model fit than does the model with predictors only at level 1 (Δ deviance = 32.47728, $df = 11$, $p < .001$). Again, however, this only addresses issues of statistical significance, not of practical significance.

The more substantive differences were found between the HLM and OLS regression analyses, which found different level-1 predictors to be significant.[1] In the OLS regression, course-related faculty interactions and intended co-curricular involvements were significant predictors of sense of belonging at the .05 level; neither predictor was significant in the HLM analysis. Although this is unsurprising, as the correct estimation of standard errors in HLM might be expected to lead to fewer significant findings, is somewhat different than Astin and Denson's (2009) finding that HLM and OLS regression did not lead to different conclusions about the significance of any predictors. It is important to note that had we chosen a more conservative alpha level of .001, as recommended by Astin and Denson (2009), the exact same predictors would have been significant in the HLM and OLS regression analyses. However, many studies in the field use an alpha level of .05. Had we only conducted the regression analysis, it is likely that we would have concluded that these predictors were in fact significant. As others have noted, our findings point to the more conservative analysis provided by HLM, which leads to a higher risk of a Type 1 error using OLS regression but a higher risk of a Type 2 error in HLM.

Beyond the simple question of significance, the three analyses above also highlight what might be gained or lost depending on decisions made about the appropriate levels of analysis. As Astin and Denson (2009) noted, OLS regression does allow the researcher to conduct a hierarchical analysis to determine how much variance can be accounted for by different groups of variables (using R^2 and ΔR^2). While there are ways to compare nested HLM models to one another, there is not an easily interpretable equivalent using HLM. OLS regression, however,

1. While the OLS regression and HLM analyses included slightly different samples (due to the handling of missing data and exclusion of 1-unit groups in HLM), these are the actual samples that would have been used had we chosen to conduct the regression or the HLM analyses. As the purpose is to compare practical scenarios, we decided to compare these two analyses despite the different sample sizes.

Table 5. Comparing results cross three models

	3-Level HLM, predictors at all 3 levels			Hierarchical OLS regression			3-Level HLM, predictors at level 1 only		
Variable	Coefficient	S.E.	*p* Value	Beta	S.E.	*p* Value	Coefficient	S.E.	*p* Value
Level 1									
Race (0 = White, 1 = Students of Color)	−.427	.068	<.001	−.421	.059	<.001	−.426	.068	<.001
Gender (0 = female, 1 = male)	.061	.056	.276	.091	.049	.063	.063	.056	.266
SES	−.007	.006	.281	.000	.006	.983	−.007	.006	.267
High school grades	−.140	.035	<.001	−.163	.029	<.001	−.140	.035	<.001
Academic transition to college	.070	.010	<.001	.070	.009	<.001	.069	.010	<.001
Social transition to college	.185	.009	<.001	.191	.009	<.001	.186	.009	<.001
Course-related faculty interactions	−.009	.011	.418	−.021	.010	.036	−.009	.011	.382
Residence hall academic climate	.094	.016	<.001	.102	.014	<.001	.100	.015	<.001
Residence hall social climate	.138	.010	<.001	.126	.009	<.001	.129	.010	<.001
Positive peer diversity interactions	.031	.006	<.001	.027	.005	<.001	.028	.006	<.001
Intended co-curricular involvement	.014	.016	.345	.029	.014	.039	.013	.015	.374
Level 2									
Mean residence hall academic climate	.047	.051	.357						
Mean residence hall social climate	−.111	.031	<.001						
Sense of belonging as an important outcome	.068	.057	.234						
Selectivity of the LLP	.101	.077	.190						
Level 3									
1,000–4,999 (vs. over 20,000)	−.659	.332	.055						
5,000–9,999 (vs. over 20,000)	−.137	.192	.482						
10,000–19,999 (vs. over 20,000)	−.075	.112	.511						
Control (1 = private not-for-profit, 0 = public)	.101	.172	.559						
Percent White	.013	.004	.004						
Selectivity	−.007	.004	.119						
Student/faculty ratio	.016	.014	.274						

does not allow for a discussion of the interaction of two predictors at different levels (in this case the interaction of individual race and LLP/university). Only through using HLM can we easily discuss the fact that while overall mean sense of belonging does vary across LLPs and universities, the difference between White students and Students of Color remains the same. Whether or not we find other significant predictors at any level, this finding has substantial implications for research and practice.

Limitations

Before moving on to implications and recommendations for future research using nested data, it is important to note a few key limitations of this study. First, the purpose of this study was to explore methodological issues involved in the use of HLM and OLS regression using real data. While this points to important practical issues that arise from the use of real data, only a pure simulation study with artificial data that can be controlled by the researcher can provide an absolute comparison across methodological approaches. For example, the analyses in this study, using real data yielded different sample sizes, where a simulation study may have held sample size constant across the three analysis techniques. Other researchers using other real datasets may have found different results in comparing the use of OLS regression and HLM, as was the case with Astin and Denson (2009). Second, although the variables chosen for inclusion in this study have a solid basis in previous literature (e.g., Johnson et al. 2007), the inclusion or exclusion of different variables was meant only to illustrate how those decisions may influence findings of significance. The findings of this study regarding the relationship between race, LLPs, and sense of belonging should be interpreted with caution, if at all.

Implications and Recommendations

Our findings point to the benefits and costs associated with decisions regarding HLM analyses in higher education. We agree with Astin and Denson (2009) that OLS regression offers benefits in stepwise modeling, but our finding of differences in the significant predictors between the HLM and OLS analyses raises concerns about Type 1 errors, particularly when a liberal p value is used ($p < .05$). Overall HLM is one strategy to account for the nesting of data, and provides the advantage of being able to explore cross-level effects. However, there are also costs to this choice in the form of an increased risk of Type II error. It may be more parsimonious to only include level-1 predictors in an HLM analysis or to use a more conservative p value or robust standard errors in OLS to account for nesting, particularly if the ICC's are small. Additionally, it may not be wise to include predictors at level-2 when the proportion of variance accounted for at the group level is particularly small. Doing this may lead to finding predictors at the group level, that while statistically significant do not have practical value. Yet, depending on the specific research questions there may also be situations in which level-2 predictors are of practical interest, even if they only account for a small amount of variance. Raudenbush and Bryk (2002) stated that even very small amounts of variation explained at the group level (e.g., as little as 2–3%) may be important depending on the research questions.

The tension between practical and statistical significance has been a topic of discussion regarding hypothesis testing for decades (O'Connell and McCoach 2008). At the center of this debate is whether a finding of statistical significance (as determined, commonly, by $p < .05$) always necessitates meaning for application of the research findings to the real world. Depending on the test that is run, statistical significance can be attributed to the format and size of the data rather than the meaning of the effect. For example, with a large enough sample size, a t-test can be significant (i.e., $p < .05$) even with a very small difference between the two group means. These two groups appear to be different in a way that is probably not due to chance or data sampling (as can be seen by a finding of $p < .05$), but the differences between the group means are not large enough to affect practice. In other words, the groups are likely different, but they are not meaningfully different. In a landmark paper, Cohen (1988) statistically derived a way to determine whether a statistical difference was practically meaningful. The resulting calculation, Cohen's D or "effect size," is frequently used in conjunction with a p-value for many statistical tests to give readers an understanding of the magnitude of an effect. Cohen's D cutoff values have also been used to determine approximate strength of effect in standardized multiple regression coefficients. Unfortunately the same kind of "effect size" that Cohen describes is infinitely more complex in HLM due to the partitioning of variance at multiple levels (Roberts and Monaco 2006). Some methodological scholars who study HLM have recommended certain effect size calculations (Hedges 2007), yet there is still not an agreed upon solution as to the best way to determine effect size in HLM (Roberts and Monaco 2006).

While there is no clear way to denote an effect size in HLM, this does not mean that scholars should shy away from discussing the practical significance of their findings, particularly in terms of the interpretation of their data. On the contrary, we recommend that higher education scholars who use HLM should be explicit in their discussion of the practical meaning of their findings, particularly in light of the proportion of variance that can be explained on each level. For example, in this study, 1.4% of the variance in sense of belonging was explained by level-2 (LLPs) and we found a variable that was statistically significant in predicting some proportion of that 1.4% of variance (mean residence hall social climate). We could decide to focus on the how to facilitate a positive social climate in order to increase sense of belonging in our discussion based on this significant finding. We could also temper this discussion with an understanding of how residence hall social climate contributes to the level-2 variance, but only minimally contributes to the overall variance in sense of belonging. In other words, LLPs do not explain almost 99% of the variance, and therefore the LLP's mean residence hall social climate explains only a small part of 1.4% of variance in students' sense of belonging. Providing this context to the significant predictor findings is important in understanding the practical meaning of the findings. We do not recommend any particular "cut off" values for the proportion of variance explained at each level necessary to denote practical significance to predictors, as that would be arbitrary and is not recommended by the HLM literature. Instead, we recommend that higher education researchers make the question of practical meaning more transparent.

There is no one right way to decide how to analyze nested data. In light of the findings of this study, we offer the following three recommendations for higher education researchers:

1. *Be transparent* The decisions that researchers make about which methods to use and how to conduct the analysis substantially influence the results, so it is important to be clear about what you do and why. A lack of transparency in the decision-making process can lead practitioners and policy makers to over- or underestimate the practical significance of the results, possibly causing them to make policy or administrative decisions that are not, in fact, supported by the data. The proportion of variance accounted for at each level should be reported in several prominent places throughout the manuscript, and should be explained clearly for those readers unfamiliar with HLM.

2. *Provide context* In addition to reporting ICCs, it is also important to contextualize the results in terms of the variance accounted for at each level. Just because the variance is significant does not mean it is meaningful, and knowing the variance explained at each level can help researchers, practitioners, and policy makers determine the utility of the findings. Variance explained at each level should be central in the discussion of the results of the unconditional model, in the discussion of the significance of the predictors at each level, and in the contextualizing of the results in the discussion and implications. Researchers should report not only the significance of predictors at each level, but also what those predictors explain in terms of the overall variance in the outcome. For example, in Pike's (2011) study, student engagement measures accounted for 51% of ICC. However, that amounted to just over 1% of the overall variance in student cognitive gains.
3. *Consider error* If choosing to use OLS regression with nested data, we recommend checking the results in HLM, even if the variance accounted for at the group level(s) is relatively small, similar to the analysis conducted by Cox et al. (2010) and Pascarella et al. (2006). Other options to account for the elevated Type I error rate in OLS regression with nested data might be to use more conservative p-values for group level variables (Park and Denson 2009), robust methods (Campbell et al. 2012), or fixed effects. Similarly, when choosing to use HLM, it may be wise to use an alpha level of .05 rather than .01, considering the elevated risk of a Type II error.

Ultimately, the choice of whether or not to use HLM and how many levels to use in the analysis depends on the data and the questions of interest. By being transparent, providing context, and considering the potential for error in both HLM and OLS regression, researchers can pull back the curtain and provide readers with a clearer picture of the research process, facilitating a more accurate interpretation of the results. As long as these fundamental research decisions are hidden by confusing and inconsistent reporting of such fundamental information as the partitioning of the variance, we run the risk of misinterpretations of our research with potentially serious implications for policy and practice.

References

Astin, A. W. (1991). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. New York: American Council on Education/Macmillan. doi 10.1007/BF00137472.

Astin, A. W., & Denson, N. (2009). Multi-campus studies of college impact: Which statistical method is appropriate? *Research in Higher Education, 50*, 354–367. doi 10.1007/s1162-009-9121-3.

Campbell, C. M., Smith, M., Dugan, J., & Komives, S. (2012). Mentors and college student leadership outcomes: The importance of position and process. *Review of Higher Education, 35*(4), 595–625.

Cheslock, J., & Rios-Aguilar, C. (2008). *Reaping (or not) the benefits of hierarchical data.* Paper presented at the annual meeting of the Association for the Study of Higher Education, Jacksonville, FL.

Cheslock, J., & Rios-Aguilar, C. (2011). Multilevel analysis in higher education research: A multidisciplinary approach. In J.C. Smart & M.B. Paulson (eds.), *Higher education: handbook of theory and research* (vol 26, pp. 85–123). Dordrecht: Kluwer Academic Publishers. doi 10.1007/978-94-007-0702-3_3.

Clark, B. R., & Trow, M. (1966). The organizational context. In T. M. Newcomb & E. K. Wilson (eds.), *College peer groups: Problems and prospects for research*. Chicago: Aldine.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Cox, B. E., McIntosh, K. L., Terenzini, P. T., Reason, R. D., & Lutovsky Quaye, B. R. (2010). Pedagogical signals of faculty approachability: Factors shaping faculty–student interaction outside the classroom. *Research in Higher Education, 51*(8), 767–788. doi 10.1007/s11162-010-9178-z.

Dedrick, R., et al. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*(1), 69–102. doi 10.3102/0034654308325581.

Di Bartolo, A., Dor, A., Fagioli, L. P., Garcia, H., Graves, D. M., Truong, K.K., & Thomas, S.L. (2011). *Quantitative Methods in Higher Education Research: An Assessment of the Field at the Turn of the 21st Century*. Paper presented at the annual meeting of the Association for the Study of Higher Education, Charlotte, NC.

Dresel, M., & Rinderman, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: A multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education, 52*(7), 717–737. doi 10.1007/s11162-011-9214-7.

Hausmann, L., Schofield, J., & Woods, R. (2007). Sense of belonging as a predictor of intentions to persist among African American and White first-year college students. *Research in Higher Education, 48*(7), 803–839. doi 10.1007/s11162-007-9052-9.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341–370. doi 10.3102/1076998606298043.

Hu, S., & Kuh, G. D. (2003). Maximizing what students get out of college: Testing a learning productivity model. *Journal of College Student Development, 44*(2), 185–203. doi 10.1353/csd.2003.0016.

Inkelas, K. K. (2007). National Study of living–learning programs 2007 Report Findings. http://drum.lib.umd.edu/bitstream/1903/10044/1/2007%20NSLLP%20Final%20Report.pdf

Ishitani, T. T. (2011). The determinants of out-migration among in-state college students in the United States. *Research in Higher Education, 52*, 107–122. doi 10.1007/s11162-010-9187-y.

Johnson, D. R., Soldner, M., Brown Leonard, J., Alvarez, P., Inkelas, K. K., Rowan-Kenyon, H., et al. (2007). Examining sense of belonging among first-year undergraduates from different racial/ethnic groups. *Journal of College Student Development, 48*(5), 525–542. doi 10.1353/csd.20076.0054.

Kim, M. M. (2001). Institutional effectiveness of women-only colleges: Cultivating students' desire to influence social conditions. *The Journal of Higher Education, 72*(3), 287–321.

Kim, M. M., & Conrad, C. F. (2006). The impact of Historically Black Colleges and Universities on the academic success of African–American students. *Research in Higher Education, 47*(4), 399–427. doi 10.1007/s11162-005-9001-4.

Kinzie, J., Thomas, A. D., Palmer, M. M., Umbach, P. D., & Kuh, G. D. (2007). Women students at coeducational and women's colleges: How do their experiences compare? *Journal of College Student Development, 48*(2), 145–165. doi 10.1353/csd.2007.0015.

Kugelmass, H., & Ready, D. D. (2011). Racial/ethnic disparities in collegiate cognitive gains: A multilevel analysis of institutional influences on learning and its equitable distribution. *Research in Higher Education, 52*(4), 323–348. doi 10.1007/s11162-010-9200-5.

Kuh, G. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter*. Washington DC: Association of American Colleges and Universities.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage Publications.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92. doi 10.1027/1614-1881.1.3.86.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell & D. B. McCoach (eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte, NC: Information Age Publishing, Inc.

Moerbeek, M. (2004). The consequences of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129–149.

Moos, R. H. (1979). Architectural, organizational, and contextual influences on living groups. *Evaluating educational environments* (pp. 67–90). San Francisco: Jossey-Bass.

Myers, C. B. (2011). Union status and faculty job satisfaction: Contemporary evidence from the 2004 National Study of Postsecondary Faculty. *The Review of Higher Education, 34*(4), 657–684. doi 10.1353/rhe.2011.0028.

Nora, A., & Cabrera, A. F. (1993). The construct validity of institutional commitment: A confirmatory factor analysis. *Research in Higher Education, 34*, 243–262.

O'Connell, A. A., & McCoach, D. B. (eds.). (2008). *Multilevel modeling of educational data*. Charlotte, NC: Information Age Publishing, Inc.

Park, J. J., & Denson, N. (2009). Attitudes and advocacy: Understanding faculty views on racial/ethnic diversity. *The Journal of Higher Education, 80*(4), 415–438. doi 10.1353/jhe.0.0054.

Pascarella, E. T., et al. (2006). Institutional selectivity and good practices in undergraduate education: How strong is the link? *The Journal of Higher Education, 77*(2), 251–285. doi 10.1353/jhe.2006.0016.

Pike, G. R., Kuh, G. D., McCormick, A. C., Ethington, C. A., & Smart, J. C. (2011). If and when money matters: The relationships among educational expenditures, student engagement and students' learning outcomes. *Research in Higher Education, 52*, 81–106. doi 10.1007/s11162-010-9183-2.

Porter, S. R. (2006). Institutional structures and student engagement. *Research in Higher Education, 47*(5), 521–558. doi 10.1007/s11162-005-9006-z.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Roberts, J. K. & Monaco, J. P. (2006, April). *Effect size measures for the two-level linear multilevel model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Smeby, J. C., & Try, S. (2005). Departmental contexts and faculty research activity in Norway. *Research in Higher Education, 46*, 593–619. doi 10.1007/s11162-004-4136-2.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research, 22*(3), 342–363.

Steinberg, M. P., Piraino, P., & Haveman, R. (2009). Access to higher education: Exploring the variation in Pell Grant prevalence among U.S. colleges and universities. *The Review of Higher Education, 32*(2), 235–270. doi 10.1353/rhe.0.0058.

Toutkoushian, R. K. (2001). Changes in demand for Public Land-Grant Universities: The use of panel datasets. *Research in Higher Education, 42*, 679–707. doi 10.1023/A:1012201511122.

Umbach, P. D., Palmer, M. M., Kuh, G. D., & Hannah, S. J. (2006). Intercollegiate athletes and effective educational practices: Winning combination or losing effort? *Research in Higher Education, 47*(6), 709–733. doi 10.1007/s11162-006-9012-9.

Volkwein, J. F., Cabrera, A. F., Lee, C., & Strauss, L. C. (2005). *Classroom Instructional Vitality: Two Competing Explanations*. Paper presented at the annual meeting of the Association for the Study of Higher Education, Philadelphia, PA.

Walsh, W. B. (1989). Person-environment interaction. In G. D. Kuh, J. P. Bean, D. Hossler, & F. K. Stage (eds.), *ASHE reader on college students* (pp. 105–113). Needham Heights: Ginn Press.